M1 Milestone Report

1. For the first milestone, our goal was to perform data collection, data preprocessing and analysis with real-time data from other people. For this, we made a Google form and asked people to fill it. However, we realized collecting enough data can be quite challenging and the number of responses might not be sufficient to train our model. Therefore, we decided to train our model with an emotion-labelled dataset from Kaggle (https://www.kaggle.com/ishantjuyal/emotions-in-text) with over 20,000 entries and test it against our collected dataset at the end.

For data preprocessing and analysis, the emotions dataset was read and the first five and last five rows from the dataset was visualized. After this, different functions were created for every text preprocessing task. These tasks include: removing punctuations, removing stopwords, removing numbers, lemmatization, converting to lowercase and tokenization. These functions were then applied on the text column of the dataset. After that, ngram extraction was done using bigram and was visualized with a horizontal bar graph and word cloud. In addition, unigrams were also visualized for comparison.

Furthermore, we have completed our untrained LSTM model for our sentiment analysis. The model's hyperparameters are not yet optimized.

2. Initially, we were planning to use Naïve Bayes based Support Vector Machine for sentiment analysis. However, we decided to use LSTM for our model and transfer learning to get better accuracy because a lot of successful models and related tasks use it.

3. Our bottlenecks include finding a suitable dataset and improving our model because our training accuracy is not up to standards. If we continue having problems increasing our model's accuracy we can further adjust the datasets by lowering the number of labels or removing inconsistent entries from the datasets.

4. Everyone worked on collecting real-time data asking people to fill in a google form.
    a. Max: Created some code for any user input responses into our program which cleans the text including removing punctuations, removing stopwords, removing numbers, removing emojis, lemmatization, converting to lowercase and tokenization.
    b. Sijan: Did data preprocessing and analysis with emotions dataset which includes doing tokenization, lemmatization, extracting dataset using bigrams, unigrams and visualizing with a horizontal bar graph and word cloud. Also created simple layout for web-page
    c. Kaleb: Implemented a sentiment analysis model with embedding and LSTM layers using PyTorch libraries. Also created the training and testing loops for the model. Currently training the model and testing

different datasets that fit our task and tuning the hyperparameters to optimize the model.

d. Ryan: Learning kivy to build the GUI for the project